A Hilbert-Huang Transform Approach Combined with PCA for Predicting a Time Series

Minjeong Park

1Department of Statistics, Seoul National University

(Received August 2011; accepted October 2011)

Abstract

A time series can be decomposed into simple components with a multiscale method. Empirical mode decomposition (EMD) is a recently invented multiscale method in Huang et al. (1998). It is natural to apply a classical prediction method such as a vector autoregressive (AR) model to the obtained simple components instead of the original time series; in addition, a prediction procedure combining a classical prediction model to EMD and Hilbert spectrum is proposed in Kim et al. (2008). In this paper, we suggest to adopt principal component analysis (PCA) to the prediction procedure that enables the efficient selection of input variables among obtained components by EMD. We discuss the utility of adopting PCA in the prediction procedure based on EMD and Hilbert spectrum and analyze the daily worm account data by the proposed PCA adopted prediction method.

Keywords: Multiscale method, empirical mode decomposition, Hilbert spectrum, Hilbert-Huang transform, principal component analysis, vector AR model.

1. Introduction

Prediction has always been an object of attention in statistics. In a classical time series analysis for predicting, a model for given data is first constructed in the time domain and then the future values are forecasted; however, most models are built under certain assumptions. As a typical example, we can consider the autoregressive (AR) model that is steadily used for analyzing a time series; however, the stationarity assumption is required. In order to alleviate such assumptions for data, we can take two different approaches. One is to construct general models having less strict assumptions, for example autoregressive integrated moving average (ARIMA) model and the other one is to decompose the data into simple components. The method that we deal with in this paper is related to simple components.

Decomposing a time series into several components can be performed by Fourier transform in the frequency domain or by wavelet transformation in the time-frequency domain. Recently, empirical mode decomposition (EMD) was invented in Huang et al. (1998), which enables to decompose a time series into several components in the time domain. Note that the decomposition process is performed

1Postdoc, Department of Statistics, Seoul National University, San 56-1 Shillim-Dong, Gwanak-Gu, Seoul 151-747, Korea. E-mail: mj.park.stat@gmail.com
in the time domain, hence we do not need an inverse transformation for reconstructing the original data. The component extracted by EMD is termed as intrinsic mode function (IMF). Through the Hilbert transform, the sum of IMFs can be represented as a generalized Fourier expansion (Huang et al., 1998) that means that the data might be decomposed into IMFs according to the spectral characteristics of the data. Based on multiscale methods such as wavelet analysis or EMD, a time series can be decomposed into several components that may be simpler than the original data.

Once the data are decomposed into simple components, it can be more feasible to analyze and predict each component by classical models. In Oh et al. (2009), a time series of Korean stock price index (KOSPI) 200 was analyzed by a multi-resolution approach that especially discussed the applications of EMD and the Hilbert spectrum in Huang et al. (1998). The applications of EMD in various fields were also briefly reviewed in Oh et al. (2009). In addition, a specific prediction procedure based on EMD and the Hilbert spectrum was proposed in Kim et al. (2008) that analyzed a real time series of cyber attacks. They suggested the following steps for prediction: 1) Obtain IMFs of a time series through the EMD process. 2) Select some IMFs based on the cumulative energy for IMFs. 3) Apply a vector AR model for the selected IMFs and polynomial regression to the EMD residue. 4) Forecast each selected IMF and the EMD residue, respectively. 5) Predict the time series by adding up all forecasting results. The prediction result by the above procedure was compared with those by exponential smoothing and the Holt-Winters model in Kim et al. (2008).

In this paper, we are interested in the usage of EMD and the Hilbert spectrum among the multiscale methods for predicting a time series. Naturally, we consider the procedure in Kim et al. (2008) for predicting a time series. Instead selecting meaningful IMFs based on the cumulative energy in Step 2); however, we propose a new method to decide input variables for the vector AR model based on the principal component analysis (PCA) in order to predict the time series more efficiently. Throughout two examples, we discuss the utility of adopting PCA in the prediction procedure based on EMD and the Hilbert spectrum. By the newly proposed method adopting PCA, we also analyze a real data set, the time series of cyber attacks in Kim et al. (2008).

This paper is organized as follows. Section 2 briefly introduces EMD and the Hilbert spectrum, and explains the procedure for predicting cyber-attacks in Kim et al. (2008). Section 3 addresses the reason to adopt PCA to the prediction procedure in Kim et al. (2008) as well as the newly proposed procedure that adopts PCA. The real worm count data that represents cyber attacks are analyzed in Section 4. Finally, Section 5 contains some concluding remarks.

2. The HHT-Based Prediction Procedure

2.1. Empirical mode decomposition

The EMD process, which was invented to analyze nonstationary and nonlinear signals in Huang et al. (1998), decomposes a signal into locally zero symmetric components, IMFs. The IMF is defined to be locally zero-symmetric in the meaning as follows: 1) The number of extrema and the number of zero crossings in the whole data set must either be equal or differ by one. 2) The mean value of the upper and lower envelopes at any point is zero. The upper (lower) envelope is defined by the spline interpolation with knots at the local maxima (minima) in the signal.

In order to extract an IMF from the given signal \( y \), the EMD process composes the following steps for sifting: 1) Identify extrema in the signal. 2) Find the upper and lower envelopes by constructing a cubic spline interpolation using maxima and minima as knots, respectively. 3) Take